CORPUS

Corpus

20 | 2020
Corpus complexes

A French text-message corpus: *88milSMS*. Synthesis and usage

Un corpus de SMS français : 88milSMS. *Synthèse et usages*

Rachel Panckhurst, Cédric Lopez and Mathieu Roche

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/corpus/4852
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Brought to you by Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture (IRSTEA)

irstea

Electronic reference
Rachel Panckhurst, Cédric Lopez and Mathieu Roche, « A French text-message corpus: *88milSMS*. Synthesis and usage », *Corpus* [Online], 20 | 2020, Online since 28 January 2020, connection on 30 January 2020. URL : http://journals.openedition.org/corpus/4852

A French text-message corpus: *88milSMS*. Synthesis and usage

Un corpus de SMS français : 88milSMS*. Synthèse et usages*

Rachel Panckhurst, Cédric Lopez and Mathieu Roche

This work was supported by the MSH-M (Maison des Sciences de l'Homme de Montpellier, France, http://www.msh-m.fr/), the DGLFLF (Délégation générale à la langue française et aux langues de France, http://www.dglflf.culture.gouv.fr/) and the CNRS (PEPS ECOMESS, HuMaIn). The SMS data described in this paper was collected within the framework of the sud4science LR (http://www.sud4science.org) project. It is part of a vast international SMS data collection project, entitled sms4science (http://www.sms4science.org), and was initiated at the CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium) in 2004. In particular, we thank Cédrick Fairon, Louise-Amélie Cougnon and Hubert Naets (CENTAL), for their support, during our project. Many thanks to our colleagues, Catherine Détrie, Claudine Moïse, Bertrand Verine. The SMS project, Sud4science LR, would never have taken place had our colleagues decided not to join us in the adventure. We are very grateful to our "Informatique et Libertés" (data protection legislation) legal advisor, Nicolas Hvoinsky, and his director, Stéphanie Delaunay (DAJI, Université Paul-Valéry Montpellier 3), who accompanied and legally advised our team throughout the project. We thank our student interns: Anthony Stifani (Master's student in Information and Communication, Université Paul-Valéry Montpellier 3), who manually analysed many of our text messages, thus allowing evaluation of the anonymization system; Pierre Accorsi and Namrata Patel (Master's students in Computer Science at the Université de Montpellier), who developed the 'Seek&Hide' software, used to anonymize the corpus; Michel Otell, Camille Lagarde-Belleville, Frédéric André and Yosra Ghliss (Master's students in Language Sciences, Université Paul-Valéry Montpellier 3) who performed the online manual anonymization with 'Seek&Hide' and verified the automatic anonymization of the corpus; Aghiles Lounes, Tarik Zaknoun Zakaria Mokrani, Reda Bestandji, Takfarinas Sider Ahmed Loudah (Master's students in Computer Science, Université de Montpellier) who worked on an automatic transcoding system. We would also like to thank the anonymous reviewers for their pertinent remarks on a previous version of our article. Any remaining mistakes are our own.

Introduction

1 The *sud4science* project (http://sud4science.org; Panckhurst *et al.*, 2013, Panckhurst *et al.* 2016b) was part of a vast international initiative, entitled *sms4science* (http://www.sms4science.org; Fairon *et al.*, 2006; Cougnon and Fairon, 2014; Cougnon, 2015). *sms4science* aimed to build a worldwide database and analyse authentic text messages in different languages — mainly French, but also Creole, German (written in Switzerland and Germany), Italian, Romansh (Dürscheid and Stark, 2011), and English (Guilbault and Drouin, 2016)[1].

2 In our previous work (Panckhurst *et al.* 2016b), we described the different methods in order to collect, to pre-process, and to publish the data of the *sud4science* project. This paper discusses and analyses the use of the *88milSMS* corpus obtained in the context of our project.

3 In this article, firstly we briefly summarise the *sud4science* data collection, ensuing processing/analysing stages, and the resulting corpus, *88milSMS*[2], through a synthesis of quotes and references to previous articles (§ 1). Secondly, we provide a state of the art on some research initiatives that use *88milSMS* in various domains and frameworks, which will enable future cross-disciplinary insight (§ 2). Then, we present other usages of the *88milSMS* corpus we identified through surveys (§ 3). Finally, we suggest future paths for textual data collection and analysis.

1. From sud4science to 88milSMS

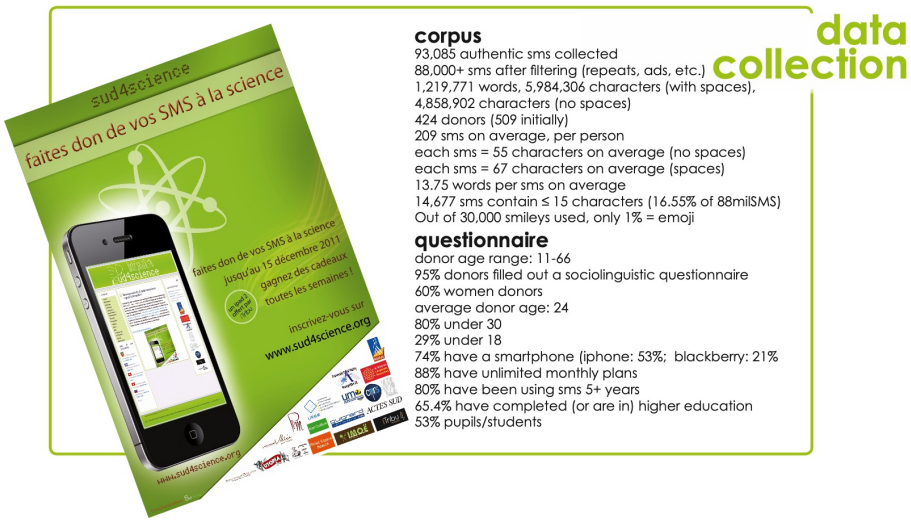
4 This section provides a schematic synthesis of both the text-message data collection project *sud4science* (http://sud4science.org), which was part of the *sms4science* international initiative (http://www.sms4science.org), and the data processing to compile the resulting *88milSMS* corpus. A more in-depth project description and analysis is provided in Panckhurst (2017: 185-235). Exhaustive references to the data-collection project and ensuing corpus can be consulted online[3].

1.1 Data collection

5 In 2011, over 88,000 authentic French text messages were collected during a 13-week period from the general public in Montpellier, France (Panckhurst *et al.* 2013, Panckhurst *et al.* 2016b) and SMS 'donors' were also invited to fill out a sociolinguistic questionnaire (Moïse 2013, Panckhurst and Moïse 2014).

6 Figure 1 provides quantitative results on the *sud4science* text-message data collection (number of SMS, characters, words, donors, smileys/emoticons, emoji) and sociolinguistic questionnaire (donor gender and age, telephone type, monthly plan, education level, etc.) (*cf.* Panckhurst *et al.* 2013: 109-111, for more detail).

Figure 1[4]. *sud4science* data collection poster and SMS / questionnaire results

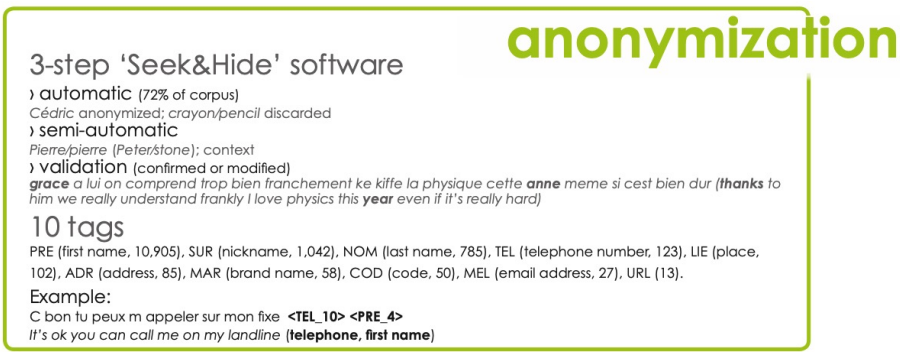


7 After the *sud4science* SMS data collection took place, there was a pre-processing phase of checking and eliminating any spurious information (including duplicates, advertisements, messages from telephone operators, etc.) (*cf.* Panckhurst *et al.* 2014b and 2014c for general explanations, details and advice).

1.2 Anonymization

8 An anonymization phase was conducted (Accorsi *et al.*, 2014, Patel *et al.*, 2013), owing to legal requirements for data-protection of private data (Ghliss and André, 2017). This involved anonymizing names, telephone numbers, places, brand names, addresses, codes, URLs (see Fig. 2 for precise tags and occurrences and § 2.2 for more detail on the semi-automatic software procedure).

Figure 2. Anonymization of *88milSMS*



1.3 Transcoding and annotation

9 Before disseminating the constructed corpus, we explored the possibilities of "transcoding" raw text messages into standardized French and linguistic "annotation". Concerning the terminology, we chose to define these terms as follows:

> "[**Transcoding**] can be defined as converting from one form of coded representation to another. This allows to discriminate between oral speech (to

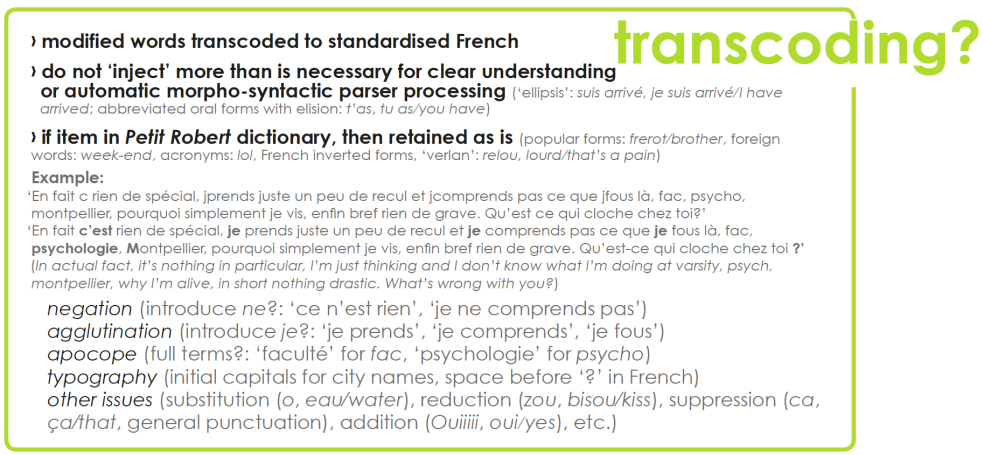
> written) 'transcription' techniques and written (to written) 'transcoding' ones, such as SMS data. From a linguistic point of view, one can also use the mainstream 'standardization', a synonym that we indeed used previously, along with 'normalization', which we prefer to use when faced with computational linguistics matters (Lopez *et al.*, 2014)." (Panckhurst, 2016: 3).
>
> "Linguistic **annotation** of SMS data for the *88milSMS* corpus [is] 'interpretative' linguistic information indicated via appropriate tags [see below] related to the difference between a 'raw' text message and its transcoded equivalent in standardized French. [We decided not to include] lemmatisation or part-of-speech (POS) tagging [...], which do indeed also correspond to other methods of linguistic annotation (based mainly on providing lexico-morpho-syntactic information)." Panckhurst (2016: 5).

10 Eight tags were chosen for linguistic annotation of *88milSMS*:

1) <TYP> (typography: punctuation, mathematical symbols, accents, numbers, hours, &, <>, (), upper and lower case, page formatting);
2) <MOD> (modification (by reduction, increase, character substitution, abbreviations, acronyms, character/phonetic repetition, interjections and onomatopoeia...): *ht* (*acheter*), *pr* (*pour*), c (*s'est, c'est, ces...*), dcd (*décider*)...);
3) <GRA> (grammar: grammatical agreement: *il viens* (*il vient*), syntax, etc.);
4) <EMO> emoji, emoticons: 🙏 :) ^^ :p ;) :d <3 :-) xd :(:/ ; 5) <ABS> (absence/ellipsis: negation, pronouns, easily identifiable missing items);
6) <LAN> (language: words borrowed from other languages, regionalisms, neologisms, French 'verlan', slang, etc.);
7) <ORT> (spelling: typing mistakes, inverted characters, etc.);
8) <DIV> (diverse: if no other tag is appropriate).

11 Lopez *et al.* (2014) discuss how "raw" anonymized text messages can be "transcoded" into "normalized" or "standardized" text messages. They use a statistical alignment method, of which the resulting prototype, entitled *AlignSMS*, enables to automatically build an SMS dictionary. Following on from the statistical *AlignSMS* method, a symbolic approach was recently proposed (Tarrade, 2017).

Figure 3. Transcoding example and related issues



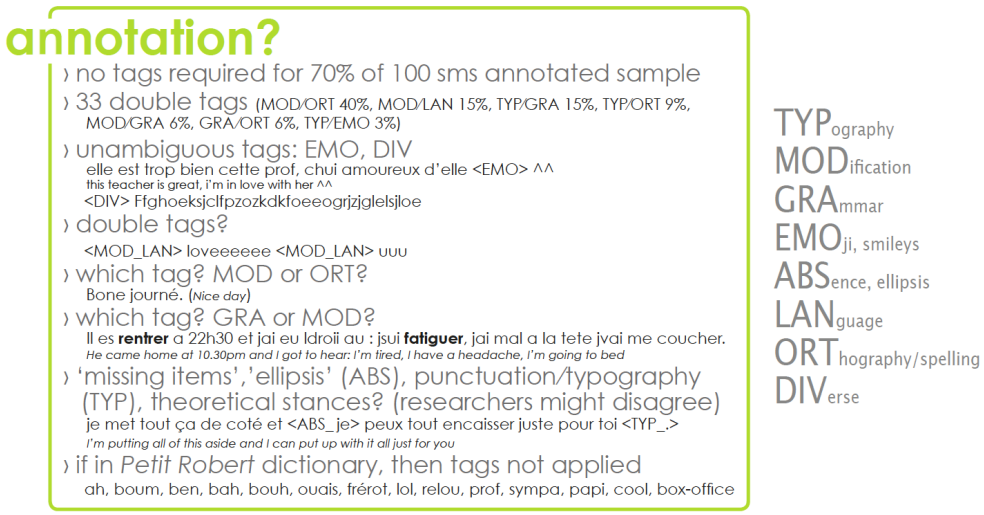
12 Panckhurst *et al.* (2016b) and Panckhurst (2016) justify exclusion of full 'transcoding' and linguistic annotation from the final processing of the *88milSMS* corpus:

> [The] (rare) choice to exclude full transcoding and tagging is a theoretical position: linguistic annotation of SMS data [...] is far from neutral. It is directly linked to an

> interpretative framework. A true consensus on how to standardize the transcoding and linguistic annotation does not exist, owing to differing/varying theoretical, (pluri)disciplinary and scientific stances. McEnery and Hardie (2012) [weigh] up the pros and cons of corpus annotation.
> [...] Mark-up initiatives should not be imposed upon researchers; it seems more relevant to let them conduct their own annotation bearing their specific scientific questioning in mind, without being trapped within a unique theoretical framework. Another alternative is that researchers may of course prefer to provide both 'raw' and tagged corpora: "Dissemination will take two different forms: one version of a corpus with the 'raw' text without any tokenization and annotation (v1), and a second version of the same corpus with the annotations (v2)." (Chanier *et al.*, 2014, p.2). For instance, Riou and Sagot (2016) present morpho-syntactic tagging of a specific corpus within the French CoMeRe corpora repository (v2), following on from a previous version without it (v1). (Panckhurst 2016: 7-8).

13 Once the corpus was fully anonymized and processed, and before dissemination, a 1,000 text-message sample was 'transcoded' into standardized French and another 100 SMS sample was linguistically 'annotated', i.e., non-standard phenomena were classified according to our annotation typology, in order to provide insight for future researchers interested in such issues, but without imposing any disciplinary related choices by the authors[5].

Figure 4. Tags and problems for SMS linguistic annotation



1.4 The *88milSMS* corpus

14 In June 2014, the finalised digital resource of 88,000 'raw' anonymized French text messages, the *88milSMS* corpus, the two samples[6] (1,000 transcoded SMS, 100 annotated SMS), and the sociolinguistic questionnaire data were made available for all to download, from the Huma-Num web service. In 2016, a TEI/XML version of the *88milSMS* corpus also became available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence on the 'Ortolang' platform[7]. The *88milSMS* corpus is the largest French SMS database ever built. Between 2014 and 2019, there have been over 780 downloads of the French from France *88milSMS* corpus from 48 countries around the world:[8]

Figure 5. Download visualization of *88milSMS* corpus per country (2014-2019)[9]

2. Research findings

15 The Montpellier *sud4science/88milSMS* project has allowed linguists, computational linguists and computer scientists, including faculty, staff and students, to collaborate on a contemporary applied research project involving authentic data. Through our project, evolving *mediated digital discourse*[10] (Panckhurst 2017) writing practices have been analysed from both language sciences and Natural Language Processing (NLP), textual data mining perspectives. The data collection, which has become a frequently consulted and cited corpus for the scientific community and beyond, provides essential concrete examples of recent societal communication.

2.1 Language Sciences perspectives

16 Below we indicate a selection of 10 general points which have emerged from our research related to evolving writing practices within the *88milSMS* **corpus:**

1. SMS-writing is very rich, innovative, creative, with no standard norm (Figure 6);[11]
2. SMSs are not highly abbreviating (the average message-reduction rate is under 10%) and character repetition/addition is sometimes used;[12]
3. Abbreviations are not solely chosen by younger generations;[13]
4. Very short messages (under 15 characters) are prevalent[14] (16.5%);
5. Scriptors' writing styles differ depending on interlocutors and contexts;
6. Interactional practices are used to maintain contact;[15]
7. Non-standard 'daily' writing similarities appear between writing practices from one century to another (e.g. 1st world war soldiers' postcards and 21st Century SMSs);[16]
8. Text-message content is often playful and emotional, sometimes ironic;[17]

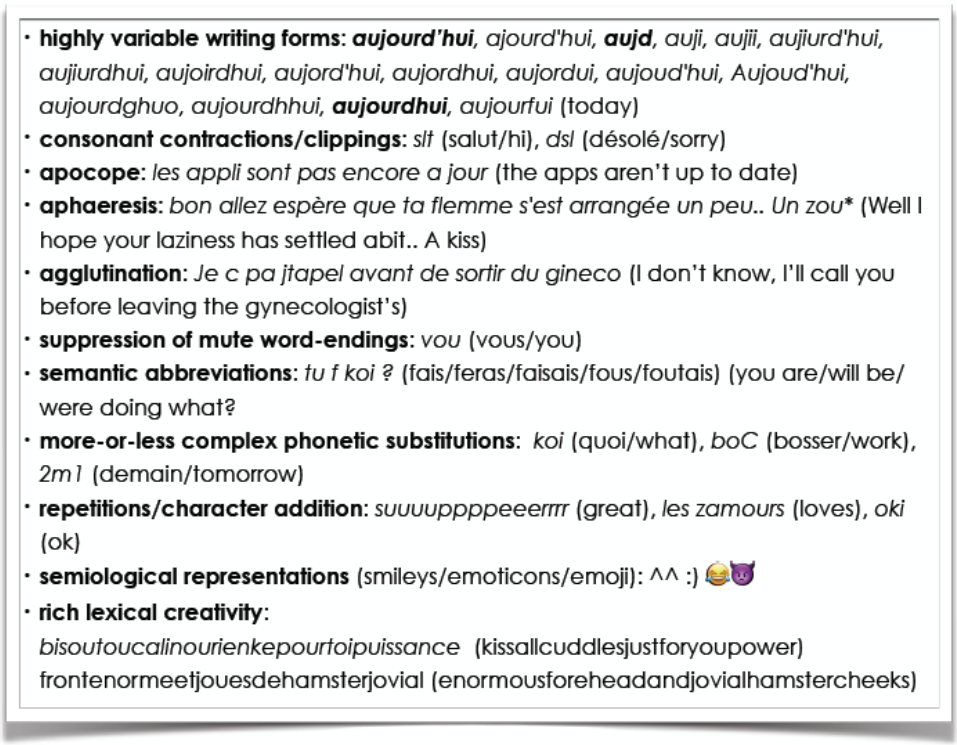
9. Textual and graphical ‘softeners’ are frequently used to decrease ambiguity, and/or aid interpretation;[18]
10. Emoticons :- ^^ <3 are common and emoji 😀🔠😍 are sometimes included.[19]

17 Other research projects include further findings on evolving writing practices: “[In *sms4science*] Cougnon (2015) showed that there was little difference between generations concerning linguistic practices; for example, there are no differences regarding words borrowed from other languages, or regionalisms. Cougnon and Draelants (2018) also showed that all generations find that respecting norms in writing conventions is very important. In terms of spelling and syntax, there are more subtle variations: verb tenses and modes are more often problematic for the young and informal question forms and negations which suppress the “ne” particle are also more apparent, which shows young people communicate in a more informal manner, but not in an incorrect one. Cougnon *et al.* (2017) compared dictations and writings over a 100-year period, and their study shows that today’s younger generations are in actual fact better at writing essays, using connectors and expressing ideas as compared to young people of yesteryear.” (Panckhurst & Cougnon, 2019). Bernicot *et al.* (2014) stipulate that texting does not impair learning traditional writing/spelling. Dürscheid and Stark (2013) study phonographic SMS writing and morpheme constancy with plurilingual examples.

18 Figure 6 shows examples of SMS-writing (see note 21 for quantative information) and we refer to the term neography in this instance:

> We define neography as writing variations which diverge from standardised language, often in a deliberate and playful manner, and are prevalent and unstable in SMS-writing. Examples are as follows: substitution (*o/eau (water), ossi/aussi (also), kikou/coucou (hi), twa/toi (you)*), addition (character repetition/punctuation : *booooooooof/bof, j’arrriiiiiiivvve !!!!!,* character addition : *les zamours*, semiological representations, emoticons/emoji, :), ^^, 😀😍), suppression (diacritic signs : *europeen/européen*; punctuation), reduction (morpho-lexical shortenings : *mdr/mort de rire (laughing out loud)*, apocopes : *ordi/ordinateur (computer)*, aphaeresis: *zou/bisou (kiss)*, double consonant suppression: *ele/elle (she),* suppression of mute word-endings: *tro/trop (too much/many)*, agglutinations: *tetrangle (strangle you)*, consonant contractions/clippings: *slt/salut (hi)*, abbreviations: *qd/quand (when)*, semantic abbreviations: *f=fais/fera(i)s/faisais ((was)doing/will do: tu f koi ? (what are you doing?))* (Panckhurst 2009, Roche *et al.* 2016).

Figure 6. Examples of neographical SMS-writing (Panckhurst 2009, Roche *et al.* 2016)



19 Sociolinguistic **questionnaire** results (Panckhurst & Moïse, 2014) indicate 5 key points[20] on why the donors text, how they use and perceive neographical SMS-writing — also related to norms and errors (Moïse 2013a, 2013b) — in the following decreasing order:

1. Cheaper or included in the monthly plan (71% of donors);
2. Quicker (69%);
3. Avoid disturbing others (50%);
4. Dislike telephoning (34%);
5. Create a close bond with friends and play with language (14%).

20 In addition to *SMS-writing, typology, semantic abbreviations* and *neography* (Roche *et al.* 2016), other aspects of MDD have been researched including: *neology* (Détrie 2017), *insults-tender words* (Détrie & Verine 2015), *forms of address* (Détrie 2014, 2015), *agreement and disagreement* (Détrie 2013, 2016), *interactional and pragmatic forms* (Panckhurst & Moïse 2011), *'isolated' and 'conversational' messages* (Panckhurst and Moïse 2012), *verbs* (Verine 2013), *emotions* (Ghliss and Verine 2016), *genres* (Verine 2015), *interjections* (Verine and Panckhurst forthcoming), *youth digital practices* (Panckhurst & Cougnon 2019).

21 A number of recent Master's and PhD dissertations allow to pursue further in-depth linguistic (André 2017, Cougnon, 2015, Guryev 2017, Morel 2017, see below) and NLP analyses (Kogkitsidou 2018, Tarrade 2017, Zenasni 2018, *cf.* §2.2.) of French SMS and instant-message writing.

22 By manual linguistic analysis of ~10,000 authentic text messages in French, from corpora collected in the *sms4science* project including *88milSMS* (Belgium, Reunion Island, Switzerland, Quebec and southern France) André (2017) shows that SMS writing

is aimed at personal appropriation of the graphic code, without orthographic standards systematically declining. He stipulates that SMS writing reveals identity, in terms of relationships to scriptors' writing and ability to adapt their discourse. The study also indicates that SMS writing can sometimes present characteristics that account for the existence of a strong link between graphic code and cognitive oralisation of a message.

23 Cougnon (2015) conducts detailed linguistic analyses of over 50,000 text messages collected from around the world within the *sms4science* project, including: language switching, neologism usage, regionalisms. She also provides descriptive and inferential statistics which give insight into modern trends of SMS-writing linked to socio-demographic variables (age, sex, education, etc.).

24 In his PhD dissertation, (Guryev, 2017) provides analysis of the syntactic variation of French interrogative structures in Swiss spontaneous electronic interaction (instant messaging, texting, WhatsApp, etc.). He postulates that under the pressure of various linguistic and non-linguistic constraints, the SMS writer chooses the particular variant which allows him/her to best achieve given communicative goals. In order to identify different types of constraints or factors that may influence the choice of variants, a multidimensional analysis model is applied which focuses simultaneously on grammatical, interactional and sociolinguistic parameters.

25 Morel (2017) analyses plurilingual practices within the Swiss *sms4science.ch* corpus (both SMS and *WhatsApp*) with French as a main language. The research focuses on three levels of regularity of plurilingual texting, i.e. (1) linguistic, (2) sociolinguistic, and (3) interactional. His PhD provides a detailed account of a pattern of plurilingualism previously unexplored.

2.2 NLP and Data Mining approaches

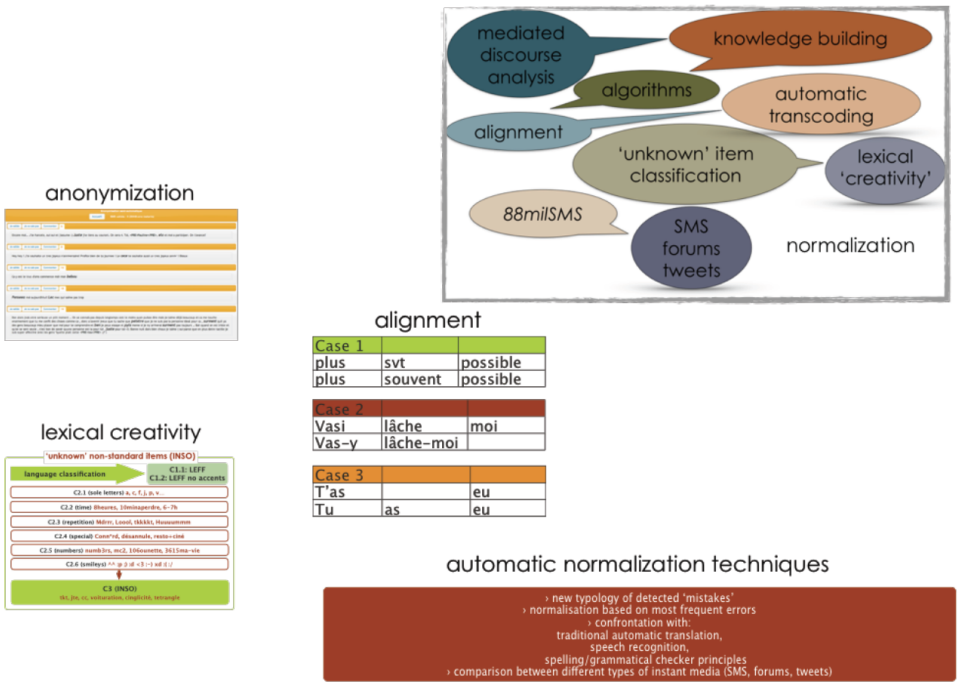
26 As specified in § 1.1., the NLP dimension of the project allowed initial processing of the data collection in particular with the 'Seek&Hide' student software for anonymization (Accorsi *et al.* 2014, Patel *et al.* 2013), and 'AlignSMS', a student alignment prototype for transcoding/normalizing French text messages (Lopez *et al*., 2014). Next, the focus was on classifying 'unknown' non-standard items (INSO) (Lopez *et al.* 2015) in text messages, thus helping to automatically identify lexical creativity[21] in *88milSMS*, which in turn may increase and improve electronic dictionary content (Figure 7).

27 Six key points summarise the computational linguistics and text-mining processing aspects of the project (see Figure 7 for a graphical representation):

1. Anonymization;
2. Alignment to transcoding;
3. INSO extraction for lexical creativity identification;
4. Normalization
5. Spatial entity recognition and extraction;
6. Sentiment analysis.

28 Real-life applications emanating from such projects could have an enormous societal impact: e.g., automatic transcoding of text messages into standardized French could be successfully incorporated into vocalizing software for those unable to consult the telephone screen (drivers, the blind, etc.).

Figure 7. Applied research



29 **Anonymization.** The *Seek&Hide* software (Accorsi *et al.*, 2014, Patel *et al.*, 2013) focuses on anonymization of identifiable information within SMS private data: first/last names, nicknames, (email) addresses, places, telephone numbers, codes, URLs, tradenames, etc. [22] First names are the main items to be hidden, but the task is difficult because different spellings can be used for a given name (*e.g. Nicolas, Nico, Nicooo, Niko, Nicoco, Nyko*). Within the framework of the *Seek&Hide* software, word-processing techniques based on a dictionary are used to label the information which needs to be anonymized. Based on such labels, the three-step semi-automatic system decides which words are to be: a) automatically anonymised, b) ignored, or c) highlighted so that the human linguist expert annotators can then process the data, via a web interface (*cf.* Fig. 8).

Figure 8. Screenshot of the 'Seek and Hide' web interface

30 **Alignment (AlignSMS).** The algorithm we proposed to align "raw" anonymized SMSs with normalized SMS is based on the pivot principle (Choudhury *et al.*, 2007) according to four steps: 1) identification of textual blocks to be aligned, 2) identification and alignment of invariant blocks (i.e. pivot blocks), 3) deducting alignments based on step 2, and 4) manual alignments of non-aligned blocks.

31 **INSO extraction for lexical creativity identification** (Lopez *et al.* 2015). Our system uses ten sequential filters in order to classify items into ten predefined categories. These categories are designed to capture all items which are not considered to be an INSO (in French *Item Non Standard Original* for *Unknown Non Standard Item*). Examples of categories are "items identifiable from lexical resources", "items without accents but identifiable in dictionaries", "items with a sole character", "hours and dates", "smileys", etc. The main idea is to capture the various items with these filters. Items that pass through all filters are considered to be INSOs (*cf.* Figure 7). This kind of resource is relevant for electronic dictionary improvement.

32 **Normalization**. Based on the *88milSMS* corpus, Tarrade (2017) develops a rule-based system using the Stanford CoreNLP architecture. These rules aim at generating normalized item candidates taking into account diacritic signs, agglutination, apocopes, consonant contractions/clippings, etc. according to a predefined typology of linguistic phenomena (Tarrade *et al.*, 2017). A score is computed for each candidate according to the kind of triggered rules and the morphosyntactic context of the item.

33 Kogkitsidou (2018) proposes a hybrid approach for automatic SMS normalization by combining fine-grained linguistic analysis based on local grammars within a machine translation model. For an information retrieval task, over the original and normalized versions of an SMS corpus, a comparison with three open source tools for name entity recognition shows that each system enhances the tagging performance over the normalized SMS.

34 **Spatial entity recognition and extraction.** Other recent research encompasses spatial recognition/extraction and sentiment analysis. (Zenasni *et al.* 2018) propose a new method combining several NLP approaches, including statistical information (*i.e.* similarity measures), lexical analysis (*i.e.* presence or absence of accents), grammatical analysis (*i.e.* part-of-speech (POS) tagging), and a text-mining approach based on n-grams of words for identifying and extracting spatial entities from the *88milSMS* corpus. The proposed methods enable to extract variations of spatial entities (*e.g. motpellier*, *montpelier*, *Montpel* are associated with *Montpellier*). Moreover, this unsupervised method has been compared with a machine learning approach in order to identify spatial entities in the *88milSMS* corpus (Lopez *et al.* 2018). It combines an approach based on Linked-Open Data for extracting rich contextual features along with standard ones that are usually included in NER systems. Both approaches (*i.e.* unsupervised and supervised) obtain comparable results.

35 **Sentiment analysis**. The work of (Khiari *et al.* 2016) presents a new opinion-mining method by combining lexical and semantic information. More precisely, the proposed approach applied to *88milSMS* gives more weight to words with a sentiment (i.e. presence of words in a dedicated dictionary) for a classification task based on three classes: positive, negative, and neutral. Moreover, the system takes into account lexical information (*e.g.* repetitions of characters) in the prediction model.

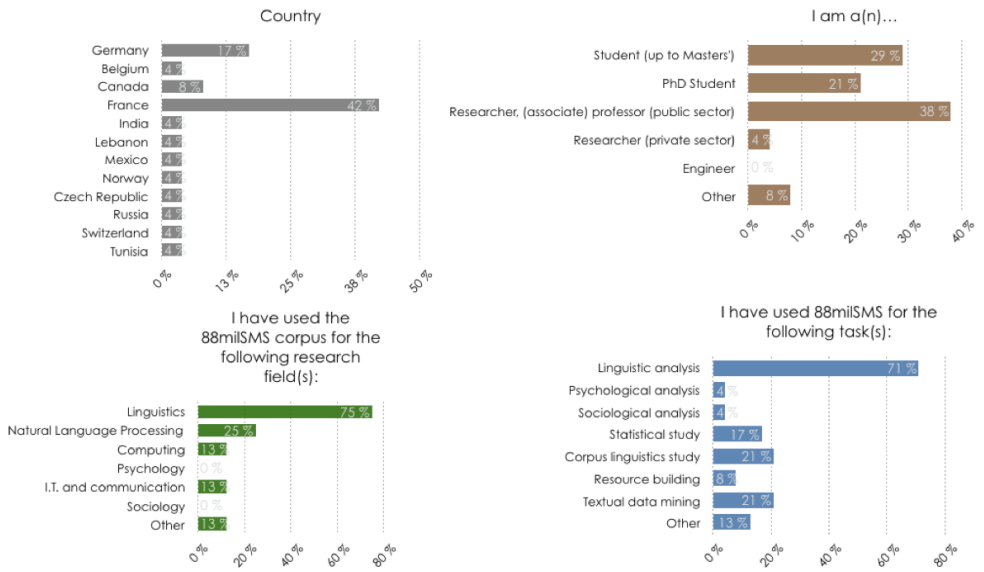
3. Surveys

36 Once the *88milSMS* corpus was uploaded to the Huma-Num platform (http://88milsms.huma-num.fr) in 2014, we gave researchers and the general public the option of signing up to a scientific newsletter.[23]

3.1 Corpus usage (2017)

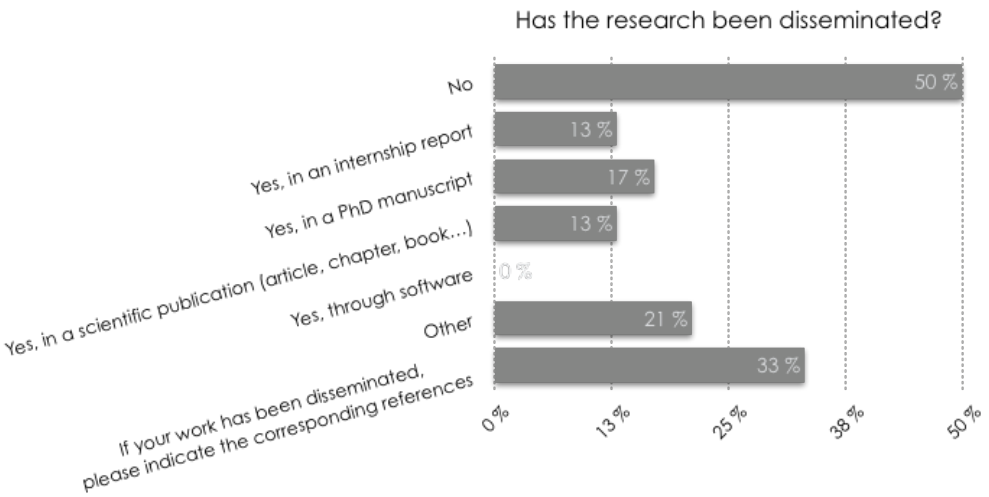
37 Three years after providing *88milSMS* for public download and dissemination, we decided to conduct a survey on usage of the corpus and asked if researchers were interested in a study day to be organised. Unfortunately, only 10% of those receiving the newsletter responded. General answers are summarised in Figure 9[24] below with a strong disciplinary tendency towards language sciences and computing including NLP, text mining and corpus linguistics research, within Europe and beyond, mainly from higher education establishments:

Figure 9. *88milSMS* usage



38 In terms of dissemination, 50% of the research cited was successfully circulated in Master's theses, PhDs, habilitations, books, articles, proceedings, etc. (Figure 10). Several colleagues and students from other disciplines contacted us in order to insert their references on our website (Kodelja *et al.* 2015, Thovex 2016).

Figure 10. Dissemination of scientific work linked to *88milSMS*



39 79% of respondents were also open to the idea of organising a workshop/conference related to the *88milSMS* corpus, with an associated publication, sometimes indicating the precise reason:

> "I'm very interested in publishing an article on analysis of the *88milSMS* corpus and showing the didactic value of text messages as a gateway to oral expression and the importance of introducing the digital register into the teaching of French as a foreign language."

3.2 Survey update (2019)

40 In March 2019, we sent an update query via the scientific newsletter to find out if colleagues had cited and/or used the *88milSMS* corpus data in their work. The survey responses received have been minimal. However, they indicate that the corpus is being used in language sciences, as is to be expected, but also in other disciplines:[25]

- Geography: Identification of place names and interpretation of variations (up-and-coming Master's 2 internship subject, 2019, IGN-Paris & Paris-Est Marne-la-Vallée University);
- Language Sciences (use *88milSMS*):

  - University courses for 2nd-year students; identifying and improving spelling mistakes (Poitiers University); discourse genres (Lorraine University);
  - recent PhD (date non-stipulated) on French as a foreign language and how to include SMS-writing in didactic situations;
  - qualitative comparative analysis between differing corpora, related to morphosyntactic French question-form usage (Guryev 2018) and interactional aspects comparing SMS and oral language (Guryev 2019);

- Psychology: digital communication and teenagers (relational, emotional romantic aspects, 12-16 year-olds, Master's 1 thesis 2019, Toulouse Jean-Jaures University).

Conclusion

41 This article provided a synthesis of the *sud4science/88milSMS* project and resulting corpus usage. In addition, this research allowed us to discover a number of new *facettes* which are not necessarily systematically investigated by academics. Also, we sometimes

needed to go beyond the institutional boundaries and this enabled to build efficient links with the Community:

1. Legal advisors: requirements for anonymizing sensitive personal SMS data;
2. Communication service: *communiqués de presse*;
3. Local firms: prize contribution during the SMS collection;
4. Media: local, national, international, written & online press, radio, TV;
5. Pluridisciplinary research;
6. Student internships, leading to student-authored publications.

42 We consider the following 4 keypoints to be fundamental for successful applied research:

1. deliver crucial research information to the general public;
2. demand that research results be factored into Ministerial reforms;
3. provide scientific expertise for devising real-life applications/software;
4. continue applied research and link academic and other institutions.

43 Also, real-life applications/software can help improve people's daily lives. Voice recognition and speech synthesis have been perfected over the decades. Our SMS research might provide insight into how electronic lexica can be modified in order to improve vocal tools used by the blind and/or those who are momentarily impeded from writing on their mobile devices.

44 Academics need to spend more time off-campus, mingling with people from other walks of life, in order to understand how their own research can become truly applied and useful for all. Links between Universities and other institutions/private enterprise are also crucial.

45 We consider SMS-writing to be one of the major creative features — an enrichment — of 21st century French written language. Analysing mediated digital discourse inevitably places researchers in the public eye. However, society often perceives contemporary writing styles in a negative fashion. As linguists and computer scientists working with NLP and text-mining, we shall continue to observe and not judge. It is our job (albeit a constant struggle) to continue to dismantle popular beliefs and convey that all written forms should be acceptable, not only standard French language. More positive ideas about technology usage and societal links need to be conveyed.

46 Recent data collections[26] (*Whatsup*, Ueberwasser and Stark 2017; *thumbs4science*, Cougnon *et al*. 2017) and future ones will continue to study evolving written language in the 21st century, i.e., investigating sociolinguistic aspects and societal impacts related to mobile technology usage and mediated digital discourse, including plurilingual and cross-cultural perspectives.

BIBLIOGRAPHY

Accorsi P., Patel N., Lopez C., Panckhurst R., Roche M. (2014). "Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques", in SMS Communication. A Linguistic Approach, (ed.) L.-A. Cougnon, C. Fairon, John Benjamins: Amsterdam/Philadelphia, 11-28.

André F. (2017). "Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française". PhD, Université Paris-Sorbonne. Jury: Cédrick Fairon, Elisabeth Stark, Rachel Panckhurst, Sylvie Plane, Gilles Siouffi (supervisor).

Androutsopoulos J. (2016). *Theorizing media, mediation and mediatization*, Sociolinguistics: Theoretical Debates (ed. N. Coupland), Cambrideg University Press.

Antoniadis G., Chabert G., Zampa V. (2011). "Alpes4science: Constitution d'un corpus de SMS réels en France métropolitain", Sherbrooke, 79th Acfas colloquium.

Barbieri F., Ronzano F. & Saggion H. (2016). "What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In Nicoletta Calzolari" (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Bernicot J., Goumi A., Bert-Erboul A., Volckaert-Legrier O. (2014). How do skilled andless-skilled spellers write text messages? A longitudinal study of sixth and seventh graders Running title: Text messages in teenagers. Journal of Computer Assisted Learning, Wiley, 30 (6),<10.1111/jcal. 12064>. <hal-01392433>

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C. R., Hriba L. Longhi J. and Seddah D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres, Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, JLCL (Journal of Language Technology and Computational Linguistics), 1-31. http://www.jlcl.org/2014_Heft2/ Heft2-2014.pdf

Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar S. and Basu A. (2007). "Investigation and modeling of the structure of texting language". *International Journal of Document Analysis and Recognition* (IJDAR), 10 (3-4), 157-174.

Cougnon L.-A. (2015). *Langage et sms. Une étude internationale des pratiques actuelles*. Louvain-la-Neuve: Presses universitaires de Louvain.

Cougnon L.-A., Draelants H. (2018). "Language Ideologies and Writing Systems in CMC: a Sociolinguistic Approach", *in* Cougnon L.-A., De Cock B. & Fairon C. (eds), *Language and the new (instant) media,* Cahiers du Cental 9, 88-99.

Cougnon L.-A., Fairon C. (eds) (2014). *SMS Communication. A linguistic approach*. Amsterdam/ Philadelphia: John Benjamins.

Cougnon L.-A., Ledegen G. (2010). "C'est écrire comme je parle. Une étude comparatiste de variétés de français dans l'écrit sms, Les voix des Français". *Modern French Identities*, 2(94), 39-57.

Cougnon L.-A., Maskens L., Roekhaut S., Fairon C. (2017). "Social media, spontaneous writing and dictation. Spelling variation", *Journal of French Language Studies* 27, 309-327.

Danesi M. (2016). *The Semiotics of Emoji: The Rise of Visual Language in the Age of the Internet*. London/ New York: Bloomsbury Publishing.

Détrie C. (2013). "Être contre et/ou être tout contre...ou comment s'accorder / se désaccorder en textotant", Colloque CJC 2013, Manifestation(s) du désaccord. De la salle de cours aux réseaux sociaux, Montpellier, 24-25 octobre 2013.

Détrie C. (2014). "(Comment) 'faire public' dans la sphère privée ? Ou le rôle des apostrophes dans les SMS", conférence invitée, Université de Lorraine, Metz, 27 mai 2014.

Détrie C. (2015). "Gentlemanminette d'amour, ma chou, colocounette et autres formes nominales d'adresse dans les SMS: de quelques spécificités liées au genre", Proceedings, Conference, "Interpréter selon les genres", April 18-20, 2013, Université Cadi Ayyad, Marrakesh, Morocco.

Détrie C. (2016). "Être contre et/ou tout contre en textotant: l'expression du consensus et du dissensus dans les SMS, entre rupture et continuum", 5th World Conference of French linguistics, F. Neveu, G. Bergounioux, M.-H. Côté, J.-M. Fournier, L. Hriba et S. Prévost (éd.), DOI : http://dx.doi.org/10.1051/shsconf/20162702004.

Détrie C. (2017). "Produire du sens en textotant: de quelques innovations lexicales, morphosyntaxiques et sémantiques dans les SMS", in Sánchez Ibáñez M., Maroto N., Torres del Rey J., De Sterck G., Linder D., García Palacios J. (eds.) *La renovación léxica en las lenguas románicas: proyectos y perspectivas.* Murcia: Editum. ISBN: 978-84-16551-75-0.

Détrie C. & Verine B. (2015). "Quand l'insulte se fait mot doux : la violence verbale dans les SMS", *in* Tuomarla U., Härmä J., Tiittula L., Sairio A., Paloheimo M. & Isosävi J. (eds) *Miscommunication and Verbal Violence. Du malentendu à la violence verbale. Miskommunikation und verbale Gewalt*. Helsinki: Société néophilologique, 59-71 (Mémoires de la Société Néophilologique de Helsinki, tome XCIII) blogs.helsinki.fi/dialog3/files/2015/07/Détrie-ja-Verine.pdf.

Drouin P., Guilbault C. (2016). "De 'Viens watcher la partie avec moi' à 'Come regarder the game with me'". Louvain-la-Neuve, Belgium. Abstracts, PLIN 2016, 12 May, http://www.plindayucl.com.

Dürscheid C., Stark, E. (2011). "sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland", *in* C. Thurlow/K. Mroczek (eds), *Digital Discourse. Language in the New Media.* Oxford: Oxford University Press, 299-320.

Dürscheid C., Stark E. (2013). "Anything goes? SMS, phonographisches Schreiben und Morphemkonstanz", in Neef M., Scherer C. *Die Schnittstelle von Morphologie und geschriebener Sprache*. Berlin: De Gruyter, 189-210.

Fairon C., Klein J.-R., Paumier S. (2006). *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*. Louvain-la-Neuve: Presses universitaires de Louvain. Manuel.CD-Rom, http://www.smspourlascience.be/

Ghliss Y., André D. (2017). "Après la collecte, l'anonymisation: enjeux éthiques et juridiques dans la constitution du corpus 88milSMS", in Wigham C. R., Ledegen G. (eds), *Corpus de communication médiée par les réseaux. Construction, structuration, analyse.* Humanités numériques. Paris: L'Harmattan, 71-84.

Ghliss Y., Verine B. (2016). Je t'aime fortttttttttt: la répétition graphémique, marqueur d'émotion dans le genre du discours SMS?, *in* Wolowska K., Krzyzanowska A. (eds) *Les Émotions et les valeurs dans la communication.* Francfort-sur-le-Main, Peter Lang.

Guilbault C., Drouin P. (2016). "Pratiques liées aux alternances de code dans un corpus anglais et français au Canada", Talk, Cercle linguistique Belge, 13 May 2016, Louvain-la-Neuve, Belgium.

Guryev A. (2017). “La forme des interrogatives dans le Corpus suisse de SMS en français: étude multidimensionnelle”, PhD, Université Paris-Sorbonne Nouvelle, en co-tutelle avec l’Université de Neuchâtel, Florence Lefeuvre, Marie-José Béguelin (supervisors).

Guryev A. (2018). “Du rôle des paramètres morphosyntaxiques dans la sélection des interrogatives”, *in* M.-J. Béguelin, A. Coveney & A. Guryev (eds) *L’Interrogative en français.* Berne: Peter Lang, 153-182.

Guryev A. (2019). “Critères de sélection des interrogatives en français: un éclairage par le biais du texto”, *in* I. Behr & F. Lefeuvre (eds) *Le Genre bref: des contraintes grammaticales, lexicales et énonciatives à son exploitation ludique et esthétique.* Frank & Timme, 109-129.

Khiari W., Bouhafs A., Roche M. (2016). "Integration of Lexical and Semantic Knowledge for Sentiment Analysis in SMS", Proceedings of LREC (International Conference on Language Resources and Evaluation), 1185-1189, Portoroz, Slovenia, 2016.

Kodelja D., Guerre M. (2015). Visualisation et analyse du réseau socio-sémantique *88milSMS*. Nantes University.

Kogkitsidou E. (2018), “Communiquer par SMS: Analyse automatique du langage et extraction de l’information véhiculée”, PhD, Université Grenoble Alpes, Jury : Antoniadis G. (supervisor), Fairon C., Kyriacopoulou T., Ledegen G., Panckhurst R., Quinard M.

Langlais P. Drouin P., Paulus A., Rompré Brodeur E., Cottin F. (2012). “Texto4Science: a Quebec French Database of Annotated Short Text Messages”, Proceedings, LREC, May, 1047-1054.

Lopez C., Bestandji R., Roche M., Panckhurst R. (2014). “Towards Electronic SMS Dictionary Construction: An Alignment-based Approach”. Proceedings, *LREC (Language Resources and Evaluation Conference)*, Reykjavik, Iceland, May 26-31, 2833-2838.

Lopez C., Roche M., Panckhurst R. (2015). “Classification des items inconnus de 88milSMS: aide à l’identification automatique de la créativité scripturale”, Travaux neuchâtelois de linguistique, 2015, 63, 71-86. https://www2.unine.ch/files/content/sites/islc/files/Tranel/63/71-86_lopez_al_corr.pdf

Lopez C., Zenasni S., Kergosien É., Partalas I., Roche M., Teisseire M., Panckhurst R. (2018). “Extracting Absolute Spatial Entities from SMS: comparing a supervised and an unsupervised approach”, in Cougnon L.-A., De Cock B., Fairon C. (eds) *Language and the new (instant) media*, coll. Cahiers du Cental, 9, Presses universitaires de Louvain, (UCL, Belgique).

Moïse C. (2013a). “Lol non tkt on ta pas oublié. Rapports à la norme et valeurs de la faute dans l’écriture Sms (projet et corpus Sud4science). Réflexions sociolinguistiques”, Plenary, Colloquium *Si j'aurais su, j'aurais pas venu ! Linguistique des formes exclues : description, genre, épistémologie*, Université Libre de Bruxelles, June 20-22.

Moïse C., (2013b). “‘Wesh trkl tkt ;) tu fou quoi ?’ La question de la norme dans l'écriture Sms : de la ‘faute’ à l’indignation normative (projet et corpus Sud4Science, http://www.sud4science.org)”, Conférence, université Laval, Québec, 18 décembre 2013.

Morel E. (2017). *Textos: assemblages hétérosémiotiques. Approche plurielle des pratiques plurilingues dans la communication par SMS et WhatsApp*. Louvain: Deboeck.

Novak, Petra Kralj, Smailović J., Sluban B. & Mozetič I. (2015). Sentiment of Emojis. *PLOS ONE* 10(12). e0144296. doi:10.1371/journal.pone.0144296.

Panckhurst R. (1997). “La communication médiatisée par ordinateur ou la communication médiée par ordinateur?”, *Terminologies nouvelles*, 17, 56-58.

Panckhurst R. (2009). “Short Message Service (SMS): typologie et problématiques futures”, in Arnavielle T. (coord.), *Polyphonies, pour Michelle Lanvin*, Université Paul-Valéry Montpellier 3, 33-52.

Panckhurst R., (2016). "A digital corpus resource of authentic anonymized French text messages: 88milSMS—What about transcoding and linguistic annotation?" *Digital Scholarship in the Humanities*. Published by Oxford University Press on behalf of EADH. http://dx.doi.org/10.1093/llc/fqw049, 11 pages.

Panckhurst R. (2017). “Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l’analyse du discours numérique médié (DNM)”, HDR defence, COMUE Université Paris-Est 30 May 2017. Jury: G. Antoniadis, C. Fairon, C. Krstev, P. Kyriacopoulou (supervisor), É. Laporte, C. Moïse, M. Roche, F. Segond.

Panckhurst R., Cougnon L.-A. (2019). “Youth Digital Practices: results from Belgian and French projects”. *TechTrends*. https://doi.org/10.1007/s11528-019-00417-y

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M. et Verine B. (2013). “Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS”. *Épistémè - revue internationale de sciences sociales appliquées*, 9 : *Des usages numériques aux pratiques scripturales électroniques*, 107-138.

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014a). "88milSMS. A corpus of authentic text messages in French", produit par l’Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l’Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014b). “Une grande collecte de SMS authentiques en français : démarche, remarques et conseils”, *Le français à l’université*, 19-03, mise en ligne le 25 septembre 2014: http://www.bulletin.auf.org/index.php?id=1875.

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014c). “Un grand corpus de SMS en français : 88milSMS”, *La lettre de l’InSHS*, 22-25, la Tribune d’Huma-Num, septembre 2014 https://www.huma-num.fr/sites/default/files/lettre_infoinshs_31_partage_experience.pdf.

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2016a). 88milSMS. A corpus of authentic text messages in French. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang : Nancy. https://hdl.handle.net/11403/comere/cmr-88milsms.

Panckhurst R., Frontini F. (forthcoming). “Emoji in text messages: Evolving interactional practices”, DeGruyter.

Panckhurst R., Moïse C., (2011). “SMS ‘conversationnels’: caractéristiques interactionnelles et pragmatiques”, 79e colloque Acfas, Sherbrooke, May 9-10, 2011.

Panckhurst R., Moïse C., (2012). “sud4science Languedoc Roussillon, collecte de SMS isolés et conversationnels. Démarche et méthode scientifiques”, communication, colloque VALS-ASLA, Lausanne, 1-3 février.

Panckhurst R., Moïse C. (2014). "French text messages. From SMS data collection to preliminary analysis", In Cougnon L.-A. and Fairon C. (eds) *SMS Communication. A Linguistic Approach.* Amsterdam/Philadelphia: John Benjamins, 141-168.

Panckhurst R., Roche M., Lopez C., Verine B., Détrie C., Moïse C., (2016b). “De la collecte à l’analyse d’un corpus de SMS authentiques: une démarche pluridisciplinaire”, in H.E.L., 38, 2, 63-82, http://www.hel-journal.org, https://www.persee.fr/issue/hel_0750-8069_2016_num_38_2.

Patel N., Accorsi P., Inkpen D., Lopez C., Roche M. (2013). “Approaches of anonymisation of an SMS corpus”, Proceedings of CICLING (Conference on Intelligent Text Processing and Computational Linguistics), LNCS, Springer Verlag, March 24-30, 2013, University of the Aegean, Samos, Greece, 77-88.

Riou S., Sagot B. (2016). Étiquetage morpho-syntaxique du corpus FAVI [corpus]. D'après Yun H. and Chanier T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. Banque de corpus CoMeRe. Ortolang.fr: Nancy. [https://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2].

Roche M., Verine B., Lopez C., Panckhurst R. (2016). “La néographie dans un grand corpus de SMS français: 88milSMS”, in J. García Palacios, G. De Sterck, D. Linder, N. Maroto, M. Sánchez Ibáñez and J. Torres del Rey (eds) *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones, Proceedings CINEO 2015, 22-24 octobre, Salamanque*. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang.: DOI: http://dx.doi.org/10.3726/978-3-631-69859-4, p. 279-302.

Steuckardt A. (2019) *Corpus 14* [Corpus]. Praxiling - UMR 5267. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/corpus14/v1.1.1.

Tarrade, L. (2017). Normalisation des messages issus de la communication électronique médiée. *Mémoire de M2R. Sciences de l'Homme et Société,* Université Grenoble Alpes 〈dumas -01666146 〉

Tarrade L., Lopez C., Panckhurst R., Antoniadis G. (2017). “Typologies pour l’annotation de textes non standard en français” *Actes du colloque TALN 2017*, 118-125, http://taln2017.cnrs.fr/actes-en-lignes/.

Thovex C. (2016). Hidden social networks analysis by semantic mining of noisy corpora IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) San Francisco, CA, USA, Aug 18-Aug 21.

Ueberwasser S. & Stark E. (2017). “What’s up, Switzerland? A corpus-based research project in a multilingual country”. *Linguistik Online* 84(5). doi:10.13092/lo.84.3849. https://bop.unibe.ch/linguistik-online/article/view/3849

Verine B. (2013). “Les verbes SMS, texto, texter et textoter dans le corpus sud4science”. https://praxiling.hypotheses.org/348

Verine B. (2015). “‘C pa 1 sms, c 1 roman!!’: le SMS est-il interprété comme un genre par ses usagers?” Proceedings, Conference, « Interpréter selon les genres », April 18-20, 2013, Université Cadi Ayyad, Marrakech, Morocco. https://halshs.archives-ouvertes.fr/hal-01317800/document.

Verine B., Panckhurst R. (forthcoming). “De la représentation de l'oral au marquage de l'interactivité: les interjections *euh* et *heu* dans le corpus 88milSMS”. Presses de l’Université de Savoie.

Zenasni, S. (2018). Extraction d'information spatiale à partir de données textuelles non-standards. Thèse de doctorat. Université de Montpellier.

Zenasni, S., Kergosien, E., Roche, M., & Teisseire, M. (2018). Spatial Information Extraction from Short Messages. Expert Systems with Applications. 95, 351-367.

NOTES

1. Several related SMS data collections took place after the initial Belgian one: Reunion Island (20,000 SMS, 2008, http://www.lareunion4science.org/; Cougnon and Ledegen, 2010), Switzerland (24,000 SMS, 2009–10, http://www.sms4science.uzh.ch; Dürscheid and Stark, 2011), Quebec (5,000 SMS, 2010, http:// www.texto4science.ca/; Langlais *et al.*, 2012), French Rhône-Alps (22,000 SMS, 2010, http://www.alpes4science.org/; Antoniadis *et al.*, 2011), and British Columbia (14,300 SMS, 2012, http://www.text4science.ca/; Drouin and Guilbault, 2016).

2. Panckhurst *et al.* 2014a (http://88milsms.huma-num.fr), Panckhurst *et al.* 2016a, TEI/XML version (https://hdl.handle.net/11403/comere/cmr-88milsms).

3. http://88milsms.huma-num.fr/references_en.html

4. Several of the following figures are extracted from previous publications: Fig. 1, 2, 3, 4 (Panckhurst 2016); Fig. 8 (Accorsi *et al.* 2014). They provide detailed discussion on the issues mentioned in these figures.

5. Some other *sms4science* corpora do indeed include both 'raw' and 'transcoded' versions of their data, but 1) the corpus size is often much smaller than *88milSMS* and 2) as specified, our position to not proceed with full transcoding, is a theoretical stance, so as to not impose specific choices which may not suit researchers from differing disciplines. Panckhurst *et al.* (2016), Panckhurst (2016) explain why annotation tags can be difficult to choose between.

6. A third extract was also later provided with a full list of the 69 (graphical) emoji (378 total occurrences) used in *88milSMS*: http://88milsms.huma-num.fr/references/emoji-88milsms.pdf " We differentiate between *emoji* (*e* 絵, "picture"; *moji* 文字, "character") 😀 🈚 ☺ 😛 😘 😷 😎 ☺ 😒

and *emoticon* ("emotion" and "icon"), the latter corresponding to mainly 'punctuation mark' usage, often requiring a 90° turn to the left :-) :) :p ;) :d :(:/ or to the right <3 in order to be interpreted, although sometimes a rotation is unnecessary, as indicated by the Japanese-influenced kaomoji emoticon ^^. [...] In *88milSMS*, around 30,000 total *emoticon* tokens and 30 different types were used." (Panckhurst & Frontini, forthcoming).

7. Panckhurst *et al.*, 2014a (http://88milsms.huma-num.fr), Panckhurst *et al.*, 2016a (https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1). We produced and submitted an XML encoding of *88milSMS*, for the Dariah initiative in 2015 (Digital Research Infrastructure for the Arts and Humanities: Dariah-fr, http://www.dariah.fr/). Our corpus was also submitted to ELRA in 2015: http://catalogue.elra.info/en-us/repository/browse/ELRA-W0082/

8. However, there may in fact have been many more downloads, since there is no mandatory form to fill out on the more recent Ortolang platform. See § 3.1 and § 3.2 for *88milSMS* corpus usage and survey information.

9. https://www.tripline.net/

10. Numerous publications on mediated digital discourse and computer-mediated communication are indicated in the bibliography at the end of this volume. This paper focusses mainly on references pertaining to *sud4science/88milSMS* and *sms4science* contexts. Panckhurst (1997) and Androutsopoulos (2016), among others, focus on differences between mediation and mediatization.

11. *Cf.* Roche *et al.* (2016) for more information.

12. Abbreviation examples : *apocope*: les **appli** sont pas encore a jour (the apps aren't up to date) — « appli » instead of « applications »*; aphaeresis*: bon allez espère que ta flemme s'est arrangée un peu.. Un **zou*** (Well I hope your laziness has settled abit.. A kiss*)* — « zou » instead of « bisou » *;*. Character repetition and addition: *suuuuppppeeerrrr* (great), les *zamours* (loves), *oki* (ok).

13. Examples: "Wesh trkl tkt ;) tu fou quoi ?" (Don't worry ;) what are you doing?) scriptor: age 12; "Ta u <PRE_5> o tel?" (Did you get hold of <NAME> on the phone?), scriptor: age 57.

14. Despite the fact that 88% of SMS donors had unlimited monthly plans in 2011, e.g., "Ok", "<3" "*Jt au resto* ^^" (I was at a restaurant ^^); semantic abbreviations, where words are reduced to

initial characters (Roche *et al*. 2016): *tu **f** koi ? (fais/feras/faisais/fous/foutais)* (you are/will be/were doing what ?). André (2017) suggests that: "very short messages remain [...] one of the characteristics of SMS-writing, of the quasi-synchronous exchanges, simulating a conversation in co-presence".

15. *88milSMS* contains solely "isolated" text messages and not "conversational" interactional ones, for legal data-collection reasons. However, we were able to infer the contact aspect from some of the very short texts (see note 16).

16. Examples: "il **à trouver** le photographe en train de les faire" (he found the photographer in the middle of doing them), soldier's postcard, 1914 (Steuckardt, Corpus 14, 2019); il **à organiser** un truc avec dès potes à là maison (he organised something with friends at home), SMS from *88milSMS*.

17. "Je te taquine <3"; "Je t'aime aussi, très fort. 😊😘🔤"; "Super !! Merci !! T'es trop gentille !!! ☹☹🔤". In the *sud4science* data collection, the number of professional messages was extremely low, indicating that more formal work-related SMSs were not donated.

18. *Cf.* Détrie & Verine (2015) for 'insults-tender words' usage, e.g., "Wesh gros ! Et bien je sais pas si je pourrai parce que jai ptetre cours, enfin jte dirai ca ce soir ^^" (Hey fatty! In actual fact I don't know if I can 'cos I might have class, I'll let you know this evening ^^).

19. 30,000 emoticons (top ten: :) ^^ :p or :P ;) :d or :D <3 :-) xd :(:/) and 378 emoji (top ten: 😊🔤😊😉😘😍😊😊🔤) are used in *88milSMS* (Panckhurst & Frontini, forthcoming). Prior research on emoji has also been conducted of course, (Danesi 2016), and emoji classifications have been proposed, including references to syntactic, semantic (Barbieri, Ronzano and Saggion 2016), semiotic, phatic and emotive/sentiment (Novak *et al*., 2015) levels).

20. Donors were invited to check one or several boxes. An open final answer was also possible: 'it leaves the choice to the receiver to answer or wait depending on the situation', 'messages which aren't important enough to phone the person', 'to re-read them', 'the pleasure of writing', 'I often write long messages, so it reduces the length without diminishing the content', 'it's amusing to write certain words in certain ways; as well as being quicker one can almost pick out who's writing the SMS with specific words and it's amusing'.

21. The most recurrent lexically creative items (Figure 6) are as follows in descending order: phonetic reduction: acronym (*lol*); graphical reduction: agglutination (*jte, jsuis, jvais*); graphical suppression: typographical elision/punctuation (*cest, weekend, Jai*); graphical reduction: consonant contractions (*Dsl, avc, Cc*); phonetic reduction: truncation (*week*); graphical substitution: diacritic signs (*méme, étre*); graphical addition: diacritic signs (*çà*); graphical addition: onomatopoeia (*Beh*); phonetic substitution with variation (*Oue*). The most frequent complex phenomena (Panckhurst 2009) are: graphical reduction : consonant contraction/ abbreviations + partial phonetic substitution: *tkt (= t'inquiète [pas])*, *pk (parce que / pourquoi);* graphical suppression: typography + graphical suppression: mute word-ending + graphical substitution with variation*: cei.*

22. *Cf.* Fig 2 for the list of 10 anonymization tags and the correlating statistics.

23. When filling out the form before downloading the corpus http://88milsms.huma-num.fr/corpus_en.html the following option can be ticked: "I wish to receive information on the scientific activity related to the corpus."

24. Percentages amount to more than 100%, as it was possible to check several boxes.

25. We also received replies from researchers (Netherlands, Switzerland) indicating they intend to use the corpus in the future, without specifying the subject area. See Kodelja *et al*. 2015, Thovex 2016 for Semantics and Computer Science references emanating from the 2017 survey.

26. *What's up, Switzerland?* (http://www.whatsup-switzerland.ch/index.php/fr), *thumbs4science* (https://acougnon.wixsite.com/thumbs4science).

ABSTRACTS

In this article, firstly we briefly summarise the *sud4science* project and data collection (http://sud4science.org), ensuing processing/analysing stages, and the resulting corpus, *88milSMS* (http://88milsms.huma-num.fr), through a synthesis of quotes and references to previous articles (§ 1). Secondly, we provide a state of the art on some research initiatives that use *88milSMS* in various domains and frameworks, which will enable future cross-disciplinary insight (§ 2). Then, we present other usages of the *88milSMS* corpus we identified through surveys (§ 3). Finally, we suggest future paths for textual data collection and analysis.

Dans cet article, nous décrivons synthétiquement le projet sud4science et la collecte de données associée (http://sud4science.org), les étapes de traitement/analyse qui en découlent et le corpus en résultant, 88milSMS (http://88milsms.huma-num.fr). Nous donnons d'abord un aperçu des travaux réalisés dans le cadre de ce projet à travers quelques citations et références (§ 1). Ensuite, nous fournissons un état de l'art sur des initiatives de recherche s'appuyant sur 88milSMS qui s'inscrivent dans des domaines et cadres de travail variés, ce qui ouvre la voie à de nouvelles perspectives interdisciplinaires (§ 2). Puis, nous présentons d'autres usages du corpus 88milSMS que nous avons identifiés via un sondage (§ 3). Enfin, nous faisons quelques propositions pour la collecte et l'analyse de données textuelles.

INDEX

Keywords: SMS, Corpus, Mediated Digital Discourse, Sociolinguistic Questionnaire, Natural Language Processing, Text Mining
Mots-clés: SMS, corpus, discours numérique médié, questionnaire sociolinguistique, traitement automatique du langage naturel, fouille de textes

AUTHORS

RACHEL PANCKHURST

Dipralang EA 739, Laboratoire de sociolinguistique, d'anthropologie des pratiques langagières et de didactique des langues-cultures, Université Paul-Valéry Montpellier 3,
rachel.panckhurst@univ-montp3.fr

CÉDRIC LOPEZ

Emvista, Montpellier, cedric.lopez@emvista.com

MATHIEU ROCHE

UMR TETIS (Univ. Montpellier, AgroParisTech, Cirad, CNRS, Irstea),
Montpellier,mathieu.roche@teledetection.fr
Cirad, Centre de coopération internationale en recherche agronomique pour le développement,
Montpellier, mathieu.roche@cirad.fr